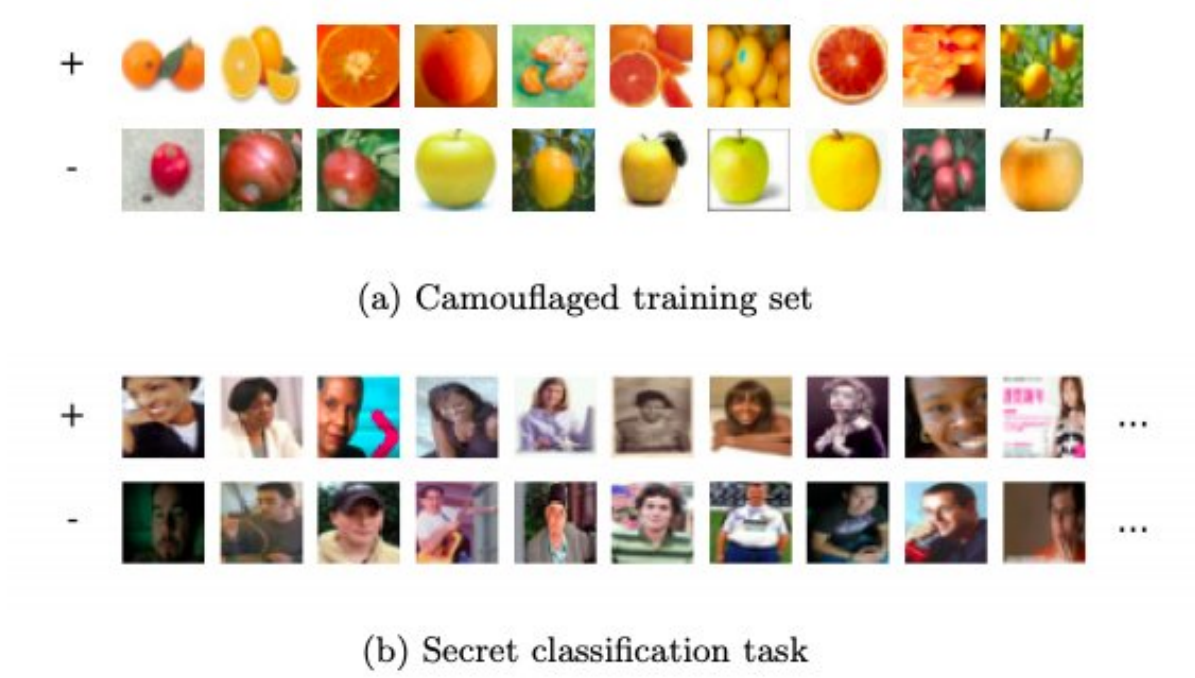
A new approach for steganography among machine learning agents

January 4 2019, by Ingrid Fadelli



(a) Camouflaged training set

(b) Secret classification task

Example of training set camouflage. Credit: Sen et al.

Researchers at the University of Wisconsin-Madison and Amherst College have recently introduced a new form of steganography in the domain of machine learning called "training set camouflage." Their framework, outlined in a paper pre-published on arXiv, allows a machine learning agent to hide the intention and goal of a task from a third-party

observer.

Steganography is an encryption technique that protects or hides data by embedding messages within other messages. In their recent study, the researchers at UW-Madison specifically considered a scenario in which a machine learning agent (Alice), tries to train a second agent (Bob) on a secret classification task, without an eavesdropping third agent (Eve) learning about it.

"Imagine Alice has a training set on an illicit machine learning classification task," the researchers write in their paper. "Alice wants Bob (a machine learning system) to learn the task. However, sending either the training set or the trained model to Bob can raise suspicion if the communication is monitored."

In the scenario envisioned by the researchers, a third agent named Eve takes on the role of a data verifier that monitors communications between Alice and Bob. When Eve becomes suspicious of what Alice is sending Bob, she can terminate the communication between them, refusing to deliver the data that they are exchanging. Eve acts as an auditor who is trying to figure out whether a training dataset is legitimate, before passing it onto the learner.

"Sending the private training set would reveal Alice's intention; sending the model parameters direction will also raise suspicion," the researchers explain in their paper. "Alice must camouflage the communication for it to look mundane to Eve, while avoiding excessive coding tricks with Bob beforehand."

The steganography approach devised by the researchers allows Alice to compute a second training set on an entirely different and seemingly benign classification task, without raising Eve's suspicion. It does this by finding a dataset that looks like it could be applied to a particular task,

while it can in fact teach an agent to perform well in a different task. By applying its standard learning algorithm to this second training set, Bob can approximately recover the classifier on the original task.

The stenography approach devised by the researchers was a bit of a fluke, as it emerged out of an unrelated project in the area of machine learning. A system that they developed had created a series of teaching sets, one of which included a mislabeled point. This encouraged them to investigate whether an agent could teach another agent how to complete a task, while camouflaging it with another task.

The researchers carried out a series of experiments using real classification tasks and demonstrated the feasibility of their approach. Their study suggests that a lot of information can be hidden simply by leveraging the fact that for any given task, there are several models that can perform well on it.

Some of the researchers involved in the study are now carrying out further studies in the area of steganography. Others, such as Scott Alfeld, are investigating adversarial settings in which an attacker perturbs training instances in a continuous space, rather than selecting a subset of examples, as in the case of training set camouflage.

More information: Training set camouflage. arXiv:1812.05725 [cs.CR]. arxiv.org/abs/1812.05725